Hyperspectral Unmixing using Transformer Network

Preetam Ghosh, Swalpa Kumar Roy, *Student Member, IEEE,* Bikram Koirala, *Member, IEEE,* Behnood Rasti, *Senior Member, IEEE,* and Paul Scheunders, *Senior Member, IEEE*

***Abstract*—Transformers have intrigued the vision research community with their state-of-the-art performance in natural language processing. With their superior performance, transformers have found their way in the field of hyperspectral image classification and achieved promising results. In this article, we harness the power of transformers to conquer the task of hyperspectral unmixing and propose a novel deep neural network-based unmixing model with transformers. A transformer network captures nonlocal feature dependencies by interactions between image patches, which are not employed in CNN models, and hereby has the ability to enhance the quality of the endmember spectra and the abundance maps. The proposed model is a combination of a convolutional autoencoder and a transformer. The hyperspectral data is encoded by the convolutional encoder. The transformer captures long-range dependencies between the representations derived from the encoder. The data are reconstructed using a convolutional decoder. We applied the proposed unmixing model to three widely used unmixing datasets, i.e., Samson, Apex, and Washington DC mall and compared it with the state-of-the-art in terms of root mean squared error and spectral angle distance. The source code for the proposed model will be made publicly available at https://github.com/preetam22n/DeepTrans-HSU.**

***Index Terms*—Hyperspectral image, unmixing, convolutional neural network, deep learning, transformer network, abundance map, endmember extraction, blind unmixing**

I. Introduction

ADVANCES in remote sensing technology improved environmental monitoring, e.g., for tracking rapid environmental changes and take precautionary actions. In particular, hyperspectral imaging (HSI) has attracted much attention in recent years. Its tasks include but are not limited to land used and land cover classification [1]–[4], forest applications [5], [6] and target detection [7] etc. In hyperspectral remote sensing, each spectral pixel might cover several pure materials on the ground due to its limited spatial resolution. The acquired spectral reflectance is then a mixture of the pure spectra (endmembers) of the materials within the pixel [8], [9]. Spectral unmixing techniques estimate the relative proportions (fractional abundances) of the endmembers within spectral pixels. The primary goal of spectral unmixing methods is to extract/estimate endmembers and their fractional abundances in each pixel by only utilizing the observed hyperspectral image. However, this often relies on the presence of a spectral library or the estimation/extraction of endmembers.

In remote sensing applications, it is generally assumed that the spectra of the pure materials are mixed linearly and several linear unmixing techniques have been developed [10]. When the endmembers of the hyperspectral image are available, the fractional abundances can be estimated by minimizing the least squared errors between the actual reflectance spectra and the ones, reconstructed by the linear model. To have a physical interpretation of the estimated fractional abundances, one must assume that no endmember can have a negative abundance. This constraint is often described as the abundance non-negativity constraint (ANC). The second constraint is the abundance sum-to-one constraint (ASC), i.e., the observed reflectance spectrum is completely composed of endmember contributions. The fully constrained least squares unmixing algorithm (FCLSU) [11] obeys both ANC and ASC. The hyperspectral pixels that follow the fully constrained linear mixing model lie on a linear simplex whose corners (vertices) are given by the endmembers. As a result, many endmember extraction algorithms have been proposed to maximize the volume enclosing simplex in the hyperspectral dataset [12]–[15]. When endmembers are not available in the hyperspectral image (no pure pixel-scenario), endmembers can be estimated by seeking the minimum volume linear simplex, which encloses the data points [16], [17]. These estimated endmembers are often denoted as virtual endmembers ([18]).

Spectral unmixing techniques that can simultaneously estimate the endmembers and the abundances are referred to as blind unmixing techniques [19]–[23]. These methods formulate the unmixing problem as a nonconvex optimization problem with respect to both endmembers and abundances. A common practice is to induce a geometrical penalty term in the fully constrained least squares method. In [24], the Euclidean distances between the estimated endmembers and the center of the hyperspectral pixels were selected to form a geometrical penalty term. In [23], the Euclidean distances between the estimated endmembers and endmembers extracted by Vertex Component Analysis (VCA) were selected for the penalty term. The total variation (TV) of all estimated endmembers was considered in [25] as a geometrical penalty. To make the technique robust to noise and outliers, in [26], a log-determinant of the estimated endmembers was considered as the geometrical penalty. Because natural images have a sparse representation in transformed domains (e.g., wavelet

This work is supported by the Research Foundation-Flanders - Project G031921N. (Corresponding author: *Bikram Koirala*)

P. Ghosh and S. K. Roy are with the Department of Computer Science and Engineering, Jalpaiguri Engineering College, West Bengal 735102, India (e-mail: pg2202@cse.jgec.ac.in; swalpa@cse.jgec.ac.in).

B. Koirala and P. Scheunders are with Imec-Visionlab, University of Antwerp (CDE) Universiteitsplein 1, B-2610 Antwerp, Belgium (e-mail: bikram.koirala@uantwerpen.be; paul.scheunders@uantwerpen.be).

B. Rasti is with Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Machine Learning Group, Chemnitzer Straße 40, 09599 Freiberg, Germany; (e-mail: b.rasti@hzdr.de).

domain) ([27]), and it is easier to remove outliers in these domains ([28]), in [29], spectral unmixing was performed in the wavelet domain. Similarly, in [27], blind unmixing was performed in the curvelet domain. The optimization equation of these methods contains a regularization parameter, which denotes the trade-off between the geometrical penalty term and the fidelity term. This parameter is data-dependent, and selecting a proper parameter for each hyperspectral image is a highly complex problem. To tackle this challenge, in [30] an automatic parameter selection technique was proposed.

If there are sufficient hyperspectral pixels on the facets of the data simplex, there exist unmixing methods to estimate virtual endmembers ([30]–[32]). When there are no data points near the facets of the data simplex, statistical methods such as [33] and [34] are a powerful alternative ([10]). When the spectral pixels are highly mixed, the estimated endmembers are not satisfactory, which leads to poor abundance maps. To deal with highly mixed scenarios, sparse unmixing techniques have been proposed [35]–[37]. These methods are often described as semi-supervised unmixing methods. Sparse unmixing utilizes a rich and well-designed library of pure spectra and applies sparse regression for the abundance estimation. A major challenge is to correct mismatches between the real reflectance spectra and the library spectra, caused by differences in the acquisition conditions of the two data types.

Due to the success of deep learning-based networks in machine learning and computer vision applications [38], [39], recently, a variety of deep neural networks has been proposed for hyperspectral unmixing ([40]). These networks are mainly based on variations of deep encoder-decoder networks. The inputs of these networks are the reflectance spectra, while the outputs are the reconstructed spectra. The encoder transforms the input spectra to the fractional abundances while the decoder transforms the abundances to the reconstructed spectra using linear layers, with the endmembers as the weights. In [41], an autoencoder-based unmixing method was proposed that improves the quality of estimated endmembers and abundance maps by incorporating spectral and spatial regularization. In [42], a two-stage network was proposed for performing blind unmixing. The first stage network estimates the endmembers and abundance maps of the input image while the second stage reconstructs the input image. In [43], autoencoders that have been used for hyperspectral unmixing are grouped into five different categories: (a) Sparse nonnegative autoencoders (a stack of nonnegative sparse autoencoders (SNSA)) [44] (b) Variational autoencoders (Deep AutoEncoder Network (DAEN) [45], Deep Generative Unmixing algorithm (DeepGUn) [46] (c) Adversarial autoencoders (Adversarial autoencoder network (AAENet)) [47]–[49] (d) Denoising autoencoders (an untied Denoising Autoencoder with Sparsity (uDAS)) [50], and (e) Convolutional autoencoders [51]–[53]. In [54], a two-stream Siamese deep network was proposed to enhance the performance of spectral unmixing. This method utilizes two subnetworks; one network learns the properties of endmembers while the other network utilizes the weights estimated by the first network to estimate abundances effectively. In [55], a multitask learning framework was utilized to refine the abundance maps estimated by the linear mixing model.

Although the advantage of incorporating the spatial information for hyperspectral unmixing has been demonstrated in the literature (especially for homogeneous regions), in SNSA, DAEN, DeepGUn, and uDAS, the spatial information is ignored. Several convolutional autoencoder-based unmixing techniques have been proposed to effectively incorporate the spatial correlation between adjacent pixels. In [56], a supervised hyperspectral unmixing method (i.e., the endmembers are assumed to be known) was proposed using a 3D convolutional autoencoder. The method referred to as unmixing using deep image prior (UnDIP) [57] utilizes endmembers extracted by a simplex volume maximization (SiVM) technique. Although several deep learning-based unmixing techniques have been specifically designed for blind unmixing, most of the methods fail when pure pixels are not available in the hyperspectral image. This is because they do not exploit the geometrical properties of the linear simplex. Recently, a minimum simplex convolutional network (MiSiCNet) [32] was proposed to incorporate both the spatial correlation between adjacent pixels and the geometrical properties of the linear simplex.

A. Contributions and Novelties

HSI, being complex in nature, pose a big challenge for Convolutional Neural Networks (CNN). As a convolution operation is limited to local features determined by the dimension of the kernel size, a significant amount of contextual information present in the original HS image is lost. Most autoencoders (AEs) are purely based on CNN networks and therefore fail to preserve a substantial portion of the original information due to the limited dimensionality of the latent space. That poses an even more significant problem in the case of HSI unmixing because the final number of endmembers is considerably lower than the initial number of spectra, causing a lot of contextual information to be lost. In real-life scenarios, a pure material is not merely limited to a local region, but can be distributed throughout the entire image. The spectral behaviour of such a pure material can vary throughout the image, due to environmental conditions. When only one spectrum of the pure material is used for unmixing, the nonlocal spatial correlation between hyperspectral pixels should be considered to improve the quality of the estimated fractional abundances. A visual transformer [58], [59] is found to be suitable for this task because it can capture nonlocal contextual feature dependencies [60]. For this task, the AE output is rearranged in terms of patches. Inspired by [58], we propose a new attention mechanism, called Multihead Self-Patch Attention to calculate the long-range dependencies between these patches. This mechanism captures long-range contextual information within the patch tokens, using a query key value system where one patch is used as the query and other patches similar to it are found from the list of keys and the best match among them is selected as the value.

This leads to better quality abundance maps and an overall better unmixing result, which in turn helps the decoder to better reconstruct the HSI. Since the weights of the decoder are used to obtain the endmember spectra, a better quality of extracted endmembers is obtained. The contribution of the proposed methodology to this end is summarized below:

- We propose a new unmixing method based on a combination of a convolutional autoencoder and a transformer. The transformer is applied to the latent space of the autoencoder to enhance the feature extraction and to ensure a better estimation of abundances and endmembers. For this, the AE output is rearranged into patches.
- Inside the transformer encoder, we propose a new attention mechanism which is referred to as Multihead Self-Patch Attention. The attention modules of the multi-head self-patch attention find the nonlocal contextual feature dependencies by determining the long-range relationship between the image patches.
- To estimate the endmembers, we apply a single convolution layer with pre-initialised weights, corresponding to the endmember spectra. These weights are learned and improved during the training of the model to obtain endmember spectra of superior quality.

The remaining of the paper is organized as follows: Section II introduces the components of the proposed method including the novel Multihead Self-Patch Attention for transformer based deep neural network-based HS image unmixing. In Section III, extensive experiments are conducted with three real datasets and one simulated dataset, and a hyperparameter sensitivity analysis and discussions are provided. Finally, comprehensive conclusions are drawn in Section IV.

II. Proposed Methodology

Let the HSI of spatial dimensions $H \times W$ with B spectral bands be denoted by $\mathbf{I} \in \mathbb{R}^{B \times H \times W}$. The HSI can be reshaped to produce the matrix $\mathbf{Y} \in \mathbb{R}^{B \times n}$, where $n = H \cdot W$ is the number of hyperspectral pixels. The endmember matrix will be denoted by $\mathbf{E} \in \mathbb{R}^{B \times R}$ where R represents the number of endmembers present in the HSI. The corresponding abundance cube (i.e., the stack of R abundance maps) is represented by $\mathbf{M} \in \mathbb{R}^{R \times H \times W}$. The abundance cube can be reshaped to produce the matrix $\mathbf{A} \in \mathbb{R}^{R \times n}$.

A. Problem formulation

In the Linear Mixing Model (LMM), the observed spectral reflectance is formulated as:

$$\mathbf{Y} = \mathbf{E}\mathbf{A} + \mathbf{N} \tag{1}$$

where $\mathbf{N} \in \mathbb{R}^{B \times n}$ is the additive noise present in $\mathbf{Y}$. Generally, three physical constraints should be satisfied: 1) the endmember matrix should be non-negative $\mathbf{E} \geq 0$; 2) ANC (Eq. (2a)); and 3) ASC (Eq. (2b)):

$$\mathbf{A} \geq 0 \tag{2a}$$

$$\mathbf{1}_R^T \mathbf{A} = \mathbf{1}_n^T \tag{2b}$$

where $\mathbf{1}_n$ indicates an n-component column vector of ones.

Since spectal unmixing is a reconstruction problem, in which abundance maps are reconstructed from the given HSI, AEs can be applied. AEs are quite capable at reconstructing and extracting information from the given inputs. In this work, the performance of an AE is complemented by the use of a transformer, to significantly improve the quality of the generated abundance maps and consequently the extracted spectral signatures of the endmembers. Fig. 1 illustrates the proposed model for deep neural network-based HSI unmixing. This figure depicts the input HS image which goes through three convolutional layers to represent the discriminative features with a fewer number of channels. After this, the HS image is broken into patches, that go through the transformer encoder consisting of a multi-head attention and feedforward layer. The output of the transformer encoder is upscaled and reshaped to match the dimension of the abundance map and a convolutional layer is used for reducing the noise. A softmax activation function is used further to enforce ASC and ANC constraints and to obtain the final abundance maps. Finally, the decoder increases the number of channels to the number of bands of the HS image by utilizing a single convolution layer whose weights are the endmembers. The components of the model are discussed in detail in subsections II-B through II-E.

B. Hyperspectral feature extraction using AE

AEs encode the input into a latent space with a lower dimensionality, learning only the salient features within the input image while avoiding unnecessary details. Owing to CNNs ability to extract high-level abstract features, using them in the encoder part of an AE provides a twofold benefit. Firstly, it heavily reduces the large number of spectral bands of a HSI and secondly, it extracts discriminative high-level features that form the base for the transformer in the next step.

The CNN applied in the encoder block of the proposed model contains three layers. Each layer progressively reduces the number of spectral bands of the HSI until C spectral bands remain. The value of C is a hyperparameter to be set. As the convolutional layer is primarily used to reduce the number of channels of the input HSI, a kernel size of 1×1 is used to keep the number of parameters low and to facilitate a faster training of the model. All three layers use a 2D convolution operation followed by a batch normalization (BN). To mitigate the vanishing gradient problem of the network, the first layer uses a dropout function. To introduce non-linearity, Leaky ReLU is used in the output of the first two layers of the AE. Table I summarizes the structure of the encoder.

TABLE I: LAYERWISE SUMMARY OF THE ENCODER BLOCK WHERE B REPRESENTS THE NUMBER OF SPECTRAL BANDS AND C IS THE NUMBER OF OUTPUT BANDS.

Layers	Composition	Kernel	Bands in	Bands out
Layer 1	Conv 2D BN Dropout Leaky ReLU	(1×1)	B	128
Layer 2	Conv 2D BN Leaky ReLU		128	64
Layers 3	Conv 2D BN		64	C

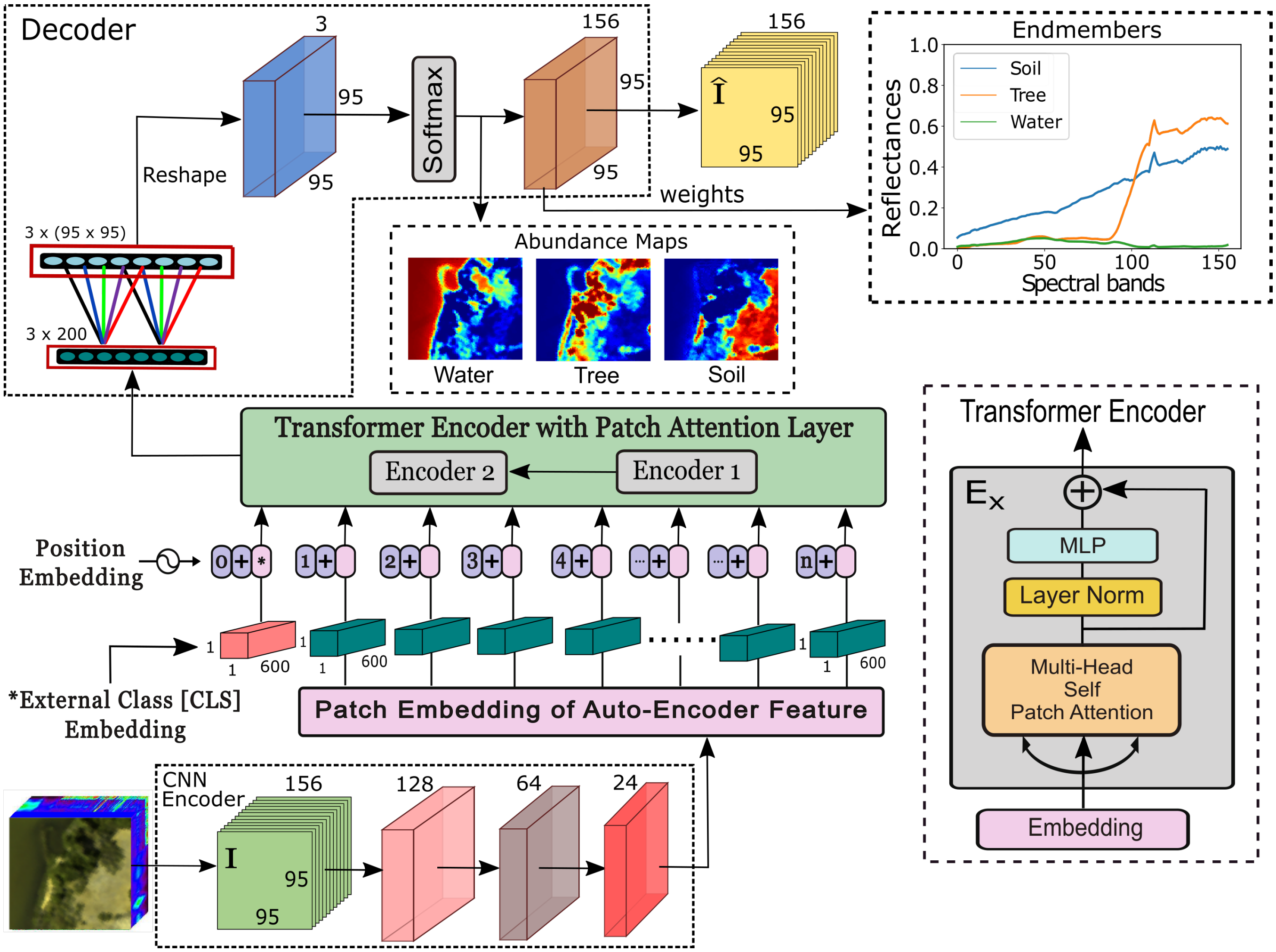


Fig. 1: Graphical representation of the proposed deep neural network-based hyperspectral unmixing model (in the figure, the Samson dataset with 95×95 hyperspectral pixels in 156 different bands is utilized to demonstrate this procedure). The input of this network is the hyperspectral image denoted by **I**. The hyperspectral image goes through three convolutional layers (CNN encoder) to represent the discriminative features with a fewer number of channels . The output of the CNN encoder is broken down into patches. The patches are reshaped into vectors and are passed through the transformer encoder consisting of a multi-head attention and multi-layer perceptron (MLP) layer. The output of the transformer encoder is upscaled and reshaped to match the dimension of the abundance map and a convolutional layer is used for reducing the noise. The softmax activation function is used further to obtain the final abundance maps. Finally, the decoder reconstructs the hyperspectral image denoted by $\hat{\mathbf{I}}$. To reconstruct the hyperspectral image, a decoder utilizes a single convolution layer whose weights are the endmembers.

In the encoder, the HSI $\mathbf{I} \in \mathbb{R}^{B\times H\times W}$ is transformed by the three consecutive layers of the encoder block into $\mathbf{I'} \in \mathbb{R}^{H\times W\times C}$:

$$\begin{aligned}
\mathbf{I_1} &= f_1(\mathbf{W_1 I} + \mathbf{U_1}) \\
\mathbf{I_2} &= f_2(\mathbf{W_2 I_1} + \mathbf{U_2}) \\
\mathbf{I_3} &= f_3(\mathbf{W_3 I_2} + \mathbf{U_3}) \\
\mathbf{I'} &= \mathbf{I_3^T}
\end{aligned} \tag{3}$$

where $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ denote the three encoder layers and $\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3}$ and $\mathbf{U_1}, \mathbf{U_2}, \mathbf{U_3}$ are the weights and biases, respectively of each layer. The superscript T denotes the matrix transpose operation.

C. Patch and Position Embeddings

To efficiently capture the long range feature dependencies, the AE output is rearranged in terms of patches. The output of the AE encoder is the cube $\mathbf{I'}$ of dimension $(H \times W \times C)$ where H, W are the spatial dimensions and C represents the reduced number of bands of the output. These features are grouped in patches $((m \cdot p) \times (n \cdot p) \times C)$ where p is the patch size and $m \cdot n$ is the total number of patches. Then the cube is reshaped to a matrix $\mathbf{X_{patch}}$ of size $((m \cdot n) \times (p \cdot p \cdot C))$ = $(N' \times D)$ where N' is the total number of patches and D is the dimension of each patch embedding. As an example, for the Samson dataset (Section III-A2), with $p = 5$ and $C = 24$, the rearrangement is given as:

$$\begin{aligned}\mathbf{I'} &= (95 \times 95 \times 24)\\ &= ((19 \cdot 5) \times (19 \cdot 5) \times 24)\\ &\rightarrow\\ \mathbf{X_{patch}} &= ((19 \cdot 19) \times (5 \cdot 5 \cdot 24))\\ &= (361 \times 600)\end{aligned}$$

In a next step, learnable class tokens $\mathbf{X_{cls}}$ of dimensions $(1 \times D)$ are defined, in which the transformer encoder will capture the long range semantic information of the patch tokens. Moreover, positional tokens $\mathbf{X_{pos}}$ of shape $(N \times D)$, with $N = N' + 1$ are generated to retain patch positional information. Rather than providing pixel and patch positional information, the positional tokens will be learned by the transformer encoder as well. Both are randomly initialized.

$\mathbf{X_{cls}}$ is appended as an extra row to the matrix $\mathbf{X_{patch}}$ and $\mathbf{X_{pos}}$ is added to the feature embedding:

$$\mathbf{X'} = (\mathbf{X_{cls}} \parallel \mathbf{X_{patch}}) + \mathbf{X_{pos}} = (\mathbf{X'_{cls}} \parallel \mathbf{X'_{patch}}) \quad (4)$$

with $\parallel$ the concatenation operation.

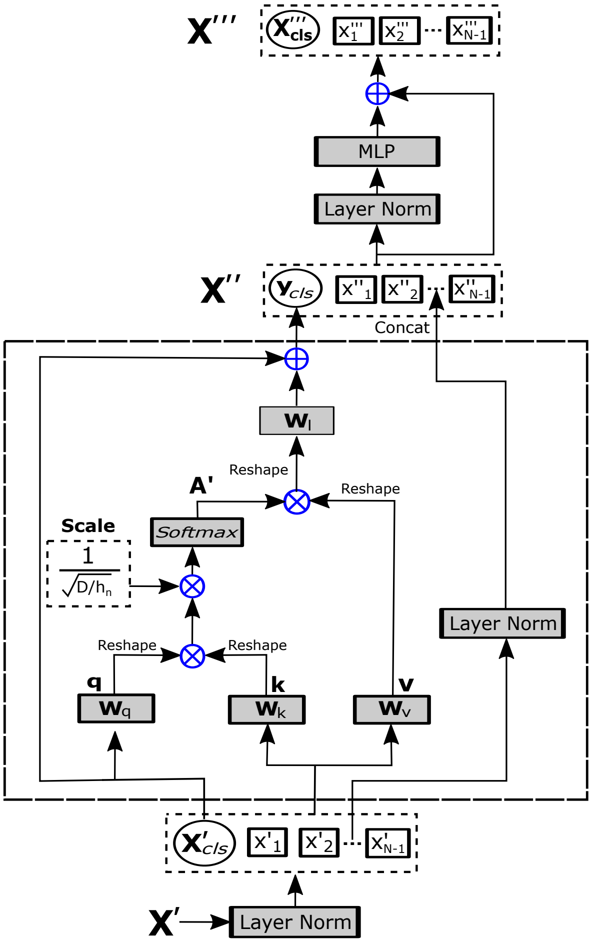


Fig. 2: Transformer Encoder with Multihead Self-Patch Attention.

D. *Transformer Encoder with Multihead Self-Patch Attention*

$\mathbf{X'}$ is the input of the next phase, which is composed of one or several transformer encoders. Each transformer encoder contains a Multihead Self-Patch Attention network [61]. The goal of this network is the exchange of information within the patch tokens to capture their long range contextual information and to feed this into the class token. The information contained in the class token is the key to improving the quality of the estimated fractional abundances. To preserve the overall patch structure, the patch tokens are appended again to the learned class token. Fig. 2 depicts the proposed Multihead Self-Patch Attention network. A detailed description of each step is given below.

Step 1: In a fist step, the overall patch matrix $\mathbf{X'}$ enters the self attention block of the transformer after going through a layer normalisation step. Attention is calculated by three linear layers. One layer works on the class token only (weight $\mathbf{W_q}$ and output $\mathbf{q}$ (queries) of size $(1 \times D)$). The other 2 layers work on the entire patch matrix (weights $\mathbf{W_k}$ and $\mathbf{W_v}$ and outputs $\mathbf{k}$ (keys) and $\mathbf{v}$ (values), both of size $(N \times D)$):

$$\mathbf{q} = \mathbf{W_q X'_{cls}}, \quad \mathbf{k} = \mathbf{W_k X'}, \quad \mathbf{v} = \mathbf{W_v X'}$$

Step 2: In the next step, the attention weight ($\mathbf{A}$) is calculated by computing the pairwise similarity between $\mathbf{q}$ and $\mathbf{k}$ and applying a softmax function:

$$\mathbf{A} = \mathtt{softmax}(\mathbf{qk^T}/\sqrt{\mathbf{D}})$$

The scaling term $(1/\sqrt{D})$ counteracts the small gradients of the softmax function. The self-patch attention ($\mathtt{PA}$) is then computed as:

$$\mathtt{PA}(\mathbf{X'}) = \mathbf{Av} \quad (5)$$

To further enhance the relationships among the different patches, self-patch attention with multiple heads is applied. For this, $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ have to reshape into matrices $\mathbf{q'}$, $\mathbf{k'}$, and $\mathbf{v'}$ of size $(h_n \times D/h_n)$, $((N \cdot h_n) \times D/h_n)$, $((N \cdot h_n) \times D/h_n)$ respectively, where h_n denotes the number of heads (attention modules). Then, the attention weight becomes:

$$\mathbf{A'} = \mathtt{softmax}(\mathbf{q'k'^T}/\sqrt{\mathbf{D/h_n}})$$

The self-patch attention with multiple heads ($\mathtt{MPA}$) is then computed as:

$$\mathtt{MPA}(\mathbf{X'}) = \mathbf{A'v'} \quad (6)$$

Step 3: The output of $\mathtt{MPA}$ is a matrix of size $(h_n \times D/h_n)$, and is then reshaped back to a matrix of size $(1 \times D)$. This matrix is further passed through a linear layer (weights $\mathbf{W_l} \in \mathbb{R}^{D \times D}$) and added up with the original class token $\mathbf{X'_{cls}}$ to obtain the class token $\mathbf{y_{cls}}$:

$$\mathbf{y_{cls}} = \mathtt{MPA}(\mathbf{X'})\mathbf{W_l} + \mathbf{X'_{cls}} \quad (7)$$

Step 4: Finally, $\mathbf{y_{cls}}$ is concatenated with the layer normalised patch tokens to obtain the output of the attention network $\mathbf{X''}$:

$$\mathbf{X''} = \mathbf{y_{cls}} \parallel \mathtt{LN}(\mathbf{X'_{patch}}) \quad (8)$$

As the output of the Multihead Self-Patch Attention network, the feature embedding $\mathbf{X''}$ is passed through a normalization layer and then fed into an Multi Layered Perceptron

Algorithm 1: Transformer Encoder with Multihead Self-Patch Attention

```
**Input:** $\mathbf{X}'$, $\mathbf{X}'_{\mathbf{cls}}$, $\mathbf{X}'_{\mathbf{patch}}$, $D$, $h_n$
**Output:** $\mathbf{X}'''_{\mathbf{cls}}$
**Multihead Self-Patch Attention (Begin)**
    **Step 1.** $\mathbf{q} = \mathbf{W_q X'_{cls}}, \quad \mathbf{k} = \mathbf{W_k X'}, \quad \mathbf{v} = \mathbf{W_v X'}$,
              $\mathbf{q}' = \text{reshape}(\mathbf{q}), \mathbf{k}' = \text{reshape}(\mathbf{k})$,
              $\mathbf{v}' = \text{reshape}(\mathbf{v})$
    **Step 2.** $\mathbf{A}' = \texttt{softmax}(\mathbf{q}'\mathbf{k}'^{\mathbf{T}}/\sqrt{\mathbf{D}/\mathbf{h_n}})$,
              $\texttt{MPA}(\mathbf{X}') = \mathbf{A}'\mathbf{v}'$ (6)
**Multihead Self-Patch Attention (End)**
**Step 3.** $\mathbf{y_{cls}} = \text{reshape}(\texttt{MPA}(\mathbf{X}'))\mathbf{W_l} + \mathbf{X}'_{\mathbf{cls}}$ (7)
**Step 4.** $\mathbf{X}'' = \mathbf{y_{cls}} \parallel \texttt{LN}(\mathbf{X}'_{\mathbf{patch}})$ (8),
          $\mathbf{X}''' = \mathbf{X}'' + \texttt{MLP}(\texttt{LN}(\mathbf{X}''))$ (9),
          $\mathbf{X}'''_{\mathbf{cls}} = \mathbf{X}'''(1,:)$
```

(MLP) block along with a residual connection to obtain the final output of the transformer encoder block (see bottom right of Fig. 1):

$$\mathbf{X}''' = \mathbf{X}'' + \texttt{MLP}(\texttt{LN}(\mathbf{X}'')) \tag{9}$$

Any number of such transformer encoders can be applied sequentially. In this work, two encoders have been applied. The output of the final block is used for further processing down the line.

The pseudo code of the Transformer Encoder with Multihead Self-Patch Attention, is shown in Algorithm 1.

E. Unmixing with decoder

The transformer produces the results $\mathbf{X}''' \in \mathbb{R}^{N \times D}$, where N is the total number of tokens and D is the dimension of each token. However, for the purpose of unmixing, only the class token $\mathbf{X}'''_{\mathbf{cls}}$ (i.e., the first row of $\mathbf{X}'''$) of size $(1 \times D)$ is considered and forwarded to the upsampling block. To do so, we reshape $\mathbf{X}'''_{\mathbf{cls}}$ to a matrix of size $R \times (D/R)$, and then upscale it to size $R \times (H \cdot W)$. Upscaling from a relatively small dimension of D/R to the dimensions $H \cdot W$ introduces noise in the final output. To solve this issue, a convolution operation with parameters $kernel\_size = (3 \times 3)$, $stride = 1$, $padding = 1$ is used. Finally, a reshaping operation is carried out to convert the output to the shape of the adundance cube $\mathbf{M}$ i.e., $(R \times H \times W)$. To ensure that the ANC and ASC constraints (Eqs. (2a) and (2b)) are satisfied, a softmax layer is used along the R dimension.

To calculate the endmembers, the abundance matrix $\mathbf{M}$ is passed through the decoder block of the AE which consists of a single convolutional layer. This convolution operation increases the number of bands in $\mathbf{M}$ from R to B, to obtain the reconstructed HSI $\widehat{\mathbf{I}}$. The weights of the convolution layer, which are initialized with the endmembers obtained from VCA, are updated throughout the learning process by the back propagation of gradients to estimate the final endmembers $\hat{\mathbf{E}} \in \mathbb{R}^{\mathbf{B} \times \mathbf{R}}$.

F. Losses and Optimization functions

In order to train the proposed model, a combination of two different losses: *Reconstruction Error (RE) loss* and *Spectral Angle Distance (SAD) loss* were applied:

$$L_{RE}(\mathbf{I}, \hat{\mathbf{I}}) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} (\hat{\mathbf{I}}_{\mathbf{ij}} - \mathbf{I}_{\mathbf{ij}})^2 \tag{10}$$

$$L_{SAD}(\mathbf{I}, \hat{\mathbf{I}}) = \frac{1}{R} \sum_{i=1}^{R} \arccos \left(\frac{\left\langle \mathbf{I}_i, \hat{\mathbf{I}}_i \right\rangle}{\|\mathbf{I}_i\|_2 \|\hat{\mathbf{I}}_i\|_2} \right) \tag{11}$$

The *RE* loss is calculated by the Mean Squared Error (MSE) objective function and helps the encoder part to learn only the essential features of the input HSI while discarding non-essential details. The *SAD* loss is a scale invariant objective function. MSE discriminates between endmembers, based on their absolute magnitude which is not desirable in case of HSI unmixing. Including SAD loss helps to counter this drawback of the MSE objective function and makes the overall model converge much faster. The total loss is calculated as the weighted sum of these two losses:

$$L = \beta L_{RE} + \gamma L_{SAD} \tag{12}$$

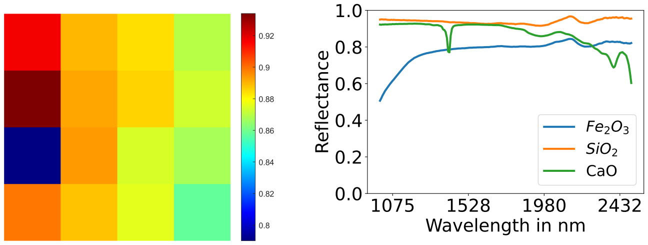
with regularization parameters β and γ.

III. Experimental Results

A. Hyperspectral Data Description

We performed experiments on four datasets. The description of the datasets are given below.

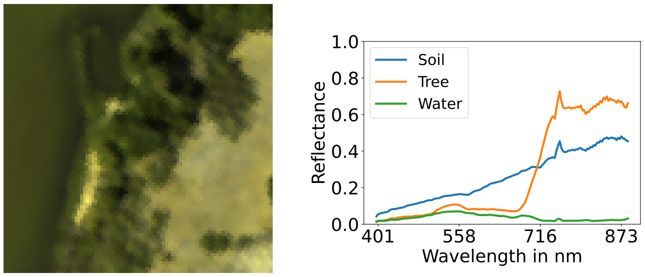
1) Simulated Dataset: A dataset of 80×80 pixels (see Fig. 3 (a)) is generated by the linear combination of three endmembers (i.e., Iron (Fe_2O_3), Silica (SiO_2), and Calcium (CaO)) (see Fig. 3(b)). Each hyperspectral pixel contains reflection values for 200 different bands covering the wavelength range [1000-2500] nm. This image contains 16 squares of 20 × 20 pixels with different ternary mixtures (see the first column of Fig. 7)



(a) (b)

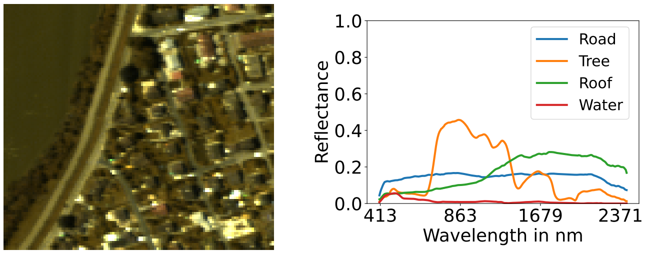
Fig. 3: The simulated image: (a) Band number 61 (1452 nm); (b) Endmembers.

2) Samson: The Samson hyperspectral dataset [62] (Fig. 4(a)) utilized in this work contains 95×95 hyperspectral pixels in 156 different bands in the wavelength range [401–889] nm. In this hyperspectral image, there are three endmembers (i.e., Soil, Tree, and Water). The ground truth endmember spectra (see Fig. 4(b)) were manually selected from the image.



(a) (b)

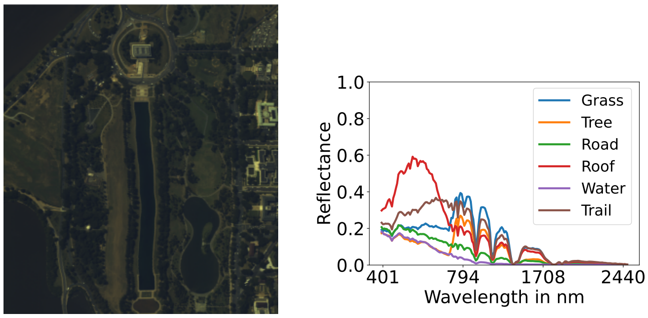
Fig. 4: Samson image: (a) True-color image (Red: 571.01 nm, Green: 539.53 nm, and Blue: 432.48 nm) (b) Endmembers.



(a) (b)

Fig. 5: Apex image: (a) True-color image (Red: 572.2 nm, Green: 532.3 nm, Blue: 426.5 nm); (b) Endmembers.

3) Apex: Fig. 5(a) shows a cropped image of the Apex dataset [63], as used in this work. This image contains 110×110 hyperspectral pixels and 285 different bands, covering the spectral channels from 413 nm to 2420 nm. There are four endmembers (i.e., Water, Tree, Road, and Roof) in this hyperspectral image. The endmember spectra shown in Fig. 5(b), were obtained from the image.



(a) (b)

Fig. 6: Washington DC Mall image: (a) True-color image (Red: 572.7 nm, Green: 530.1 nm, Blue: 425.0 nm); (b) Endmembers.

4) Washington DC Mall: This hyperspectral image is acquired over the Washington DC Mall using the HYDICE sensor [1]. Fig. 6 (a) shows the cropped data used in this paper that contains 290 × 290 pixels, in 191 different bands ranging from the wavelength 400 nm to 2400 nm. There are six endmembers (i.e., Grass, Tree, Roof, Road, Water, and Trail) in this hyperspectral image. The spectra of these materials (see Fig. 6(b)) were collected from this hyperspectral image.

The ground truth fractional abundances of all three real datasets were produced by applying FCLSU. Visually, the estimated abundance maps represent the scene (see e.g., Fig. 4(a) for the RGB image and the first column of Fig. 9 for the ground truth abundance maps of the Samson dataset). However, the ground truth abundance maps produced in this way may deviate from the real abundance maps.

B. *Experimental Setup*

The performance of the proposed model is evaluated and compared to six different unmixing techniques from different categories: **Geometrical unmixing** method `FCLSU` [11] using VCA [13] for endmember extraction, **Geometrical and blind unmixing** method `NMF-QMV` [30], **Sparse unmixing** method Collaborative LASSO (`Collab`) [64] and **Deep neural network-based unmixing** methods `uDAS` [50], `UnDIP` [57], and `CyCUNet` [53].

C. *Hyperparameters*

In deep neural network-based unmixing models, the produced results are typically dependent on the hyperparameter settings. The transformer depends on the patch size p and the transformer input dimensionality C. Moreover, regularization parameters β and γ have to be set. The model is trained during a number of epochs at a certain learning rate. Typically, an initial learning rate was set and was then gradually reduced by 20% after every 15 epochs, except for the simulated dataset, where it was reduced by 20% after every 30 epochs. Finally, a weight decay rate was incorporated in the optimization function to keep the losses in check.

Table II shows the hyperparameters chosen for training the proposed model.

TABLE II: HYPERPARAMETERS USED FOR TRAINING THE PROPOSED MODEL.

Hyperparameters	Simulated	Samson	Apex	WDC Mall
p	(8×8)	(5×5)	(5×5)	(10×10)
C	12	24	32	24
β	1×10^4	5×10^3	5×10^3	5×10^3
γ	5×10^{-2}	3×10^{-2}	5×10^{-2}	1×10^{-4}
Epoch	1000	200	200	150
Learning rate	4×10^{-3}	6×10^{-3}	9×10^{-3}	6×10^{-3}
Weight decay	5×10^{-5}	4×10^{-5}	4×10^{-5}	3×10^{-5}

D. *Quantitative Performance Measures*

Quantitative results are provided by the root mean squared error (RMSE) between the estimated and ground truth abundance fractions:

$$\mathrm{RMSE}(\mathbf{M},\hat{\mathbf{M}}) = \sqrt{\frac{1}{RHW}\sum_{k=1}^{R}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\hat{\mathbf{M}}_{kij}-\mathbf{M}_{kij}\right)^2} \tag{13}$$

[1] https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

and the spectral angle distance (SAD) in degree between the estimated and ground truth endmembers:

$$\mathrm{SAD}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{R} \sum_{i=1}^{R} \arccos \left(\frac{\langle \mathbf{s}_{(i)}, \hat{\mathbf{s}}_{(i)} \rangle}{\left\| \mathbf{s}_{(i)} \right\|_2 \left\| \hat{\mathbf{s}}_{(i)} \right\|_2} \right), \tag{14}$$

where $\langle . \rangle$ denotes the inner product and $\mathbf{s}_{(i)}$ indicates the ith column of the ground truth endmembers matrix $\mathbf{S}$.

E. Unmixing Experiments: Quantitative Results

Simulated Dataset: Quantitative results on the simulated dataset can be found in Tables III and IV. The obtained RMSE with the proposed method is comparable to the best methods and the model is consistent across different levels of noise, managing to hold the second best place for all tested levels of noise. The extracted endmembers also closely resemble the ground truth endmembers with comparable performance to the competing methods.

Samson Dataset: Quantitative results on the Samson dataset can be found in Tables V and VI. The results confirm that the proposed model outperforms the other techniques in terms of both abundance and endmember estimation with a mean RMSE of 0.0783 showing a 48.02% improvement to the next best method and a mean SAD value of 0.0608 which amounts to a 35.93% improvement.

Apex Dataset: Quantitative results on the Apex dataset can be found in Tables VII and VIII. The endmember "Road" of the Apex dataset is found to be quite a challenge for the other methods, while the proposed method estimates this endmember satisfactorily. The proposed method outperforms the other unmixing techniques with a mean RMSE value of 0.1264 and a mean SAD value of 0.0867. Additionally, it provides the best endmember estimation for Road and Water in terms of SAD.

Washington DC Mall Dataset: Quantitative results on the Washington DC Mall dataset can be found in Tables IX and X. Among all the considered datasets, the similarity between the spectral signatures of its six endmembers provides the greatest challenge. The "Tree" and "Grass" endmembers have almost identical spectral signatures, and most methods struggle to find the difference. The proposed model successfully separated these two endmembers due to its ability to find long-range dependencies among the image patches, thus leading to a RMSE value of 0.1661 and 0.0963 for the "Grass" and "Tree" endmembers, respectively. In terms of overall RMSE and SAD, the proposed model outperforms the closest competitor method by 43.71% and 52.11% respectively.

Overall observations: From Tables V, VII and IX, one can conclude that the overall performance of the proposed method beats the other competing methods by a significant margin in terms of RMSE. `Collab`, `FCLSU`, and `NMF-QMV` also shows decent performance on the Samson, Apex and WDC Mall datasets, but their performance is not consistent across the datasets. `UnDIP` and `uDAS` were unable to beat any of the methods for any given class; however, their performance was consistent throughout the different datasets used in the experiments. `CyCU` produced mixed results within a given dataset, with good performance on particular endmembers and significantly worse on other endmembers. For the simulated dataset in Table III, the model fails to secure the top spot, but the results are second best among all the other methods and only miss the top by a slight fraction.

Tables VI, VIII and X make it clear that obtaining good spectral signatures for the endmembers is more difficult than producing a good abundance map. The proposed model considerably outperforms all the other competing methods. On the Apex and WDC Mall datasets, the proposed model obtains SAD values of 0.0867 and 0.1537, respectively, about half of the next best method. For the simulated dataset (see Table IV), `Collab` and `uDAS` were the most consistent throughout the different noise levels.

It is worth mentioning that a good SAD value does not necessarily guarantee good abundance maps, because SAD removes the norm of the endmember spectra. In other words, it ignores endmember scaling factors, caused by multiple reflections of the light and continuously variable illumination conditions in practical situations. However, such scaling factors can considerably affect the abundance estimation. As the proposed method provides the best results in both SAD and RMSE, one can conclude that it overcomes the mentioned problem at least to some degree.

F. Visual Analysis of Abundance Maps and Endmembers

The abundance maps and spectral signatures of the endmembers provide a way to visually compare the generated results of the different unmixing methods. Figs. 7, 9, 11, and 13 show the abundance maps obtained from the various competing methods. It can be inferred that the abundance maps obtained from the proposed method are visually most similar to the ground truth abundance maps. For the simulated dataset, abundance maps produced by all methods, but `CyCU`, are visually similar (see Fig. 7) to the ground truth abundance maps. The methods `UnDIP` and `uDAS` fail to properly represent the endmember "Water" across all the real experimental datasets. Decent results are obtained by the methods `Collab`, `FCLSU`, and `NMF-QMV`, but their performance suffers from inconsistencies in RMSE values from one endmember to another. This causes the models to lose in terms of overall performance, even if they manage to obtain good results on a particular endmember. For example, none of the competing methods was able to correctly produce the "Road" endmember in the Apex dataset. It is interesting to observe that the abundance map of roads (see Fig. 11, Apex dataset), estimated by the proposed method seems to be more realistic than our produced ground truth abundance map. The success of the proposed model on this endmember can be attributed to the ability of the transformer encoder block with self-patch attention to find the long distance feature dependencies, which are otherwise lacking in the abundance maps obtained from the output of the convolutional network.

Figs. 8, 10, 12 and 14 depict the extracted endmembers. It was observed that the methods using `VCA` as initialization could not further improve the `VCA` extracted endmembers by much, leading to higher values of SAD later on. The proposed method however is also initialized by `VCA` analysis,

TABLE III: RMSE (SIMULATED DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	FCLSU	NMF-QMV	UnDIP	uDAS	Proposed
20dB	0.13887±0.00008	**0.06201±0.00100**	0.07083±0.00719	0.07905±0.00313	0.07310±0.00188	0.09769±0.00215	0.06744±0.00277
30dB	0.14012±0.00004	**0.02251±0.00056**	0.02543±0.00235	0.03108±0.00132	0.02615±0.00028	0.03689±0.00073	0.02432±0.00190
40dB	0.14052±0.00001	0.00911±0.00017	**0.00792±0.00031**	0.01050±0.00027	0.00880±0.00006	0.01233±0.00033	0.00850±0.00057
50dB	0.14051±0.00001	0.00578±0.00004	**0.00276±0.00028**	0.00338±0.00011	0.00458±0.00004	0.00404±0.00023	0.00292±0.00027

TABLE IV: SAD (SIMULATED DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	VCA	NMF-QMV	SiVM	uDAS	Proposed
20dB	0.04183±0.00251	**0.01500±0.00065**	0.02347±0.00219	0.02626±0.00154	0.02243±0.00094	0.01892±0.00137	0.02115±0.00110
30dB	0.04466±0.00617	0.00497±0.00020	0.00712±0.00053	0.00783±0.00031	0.00584±0.00014	**0.00392±0.00046**	0.00600±0.00070
40dB	0.04150±0.00725	0.00158±0.00006	0.00222±0.00011	0.00240±0.00008	0.00180±0.00005	**0.00065±0.00008**	0.00158±0.00010
50dB	0.03827±0.00658	**0.00051±0.00003**	0.00074±0.00006	0.00076±0.00003	0.00105±0.00005	0.00093±0.00003	0.00060±0.00010

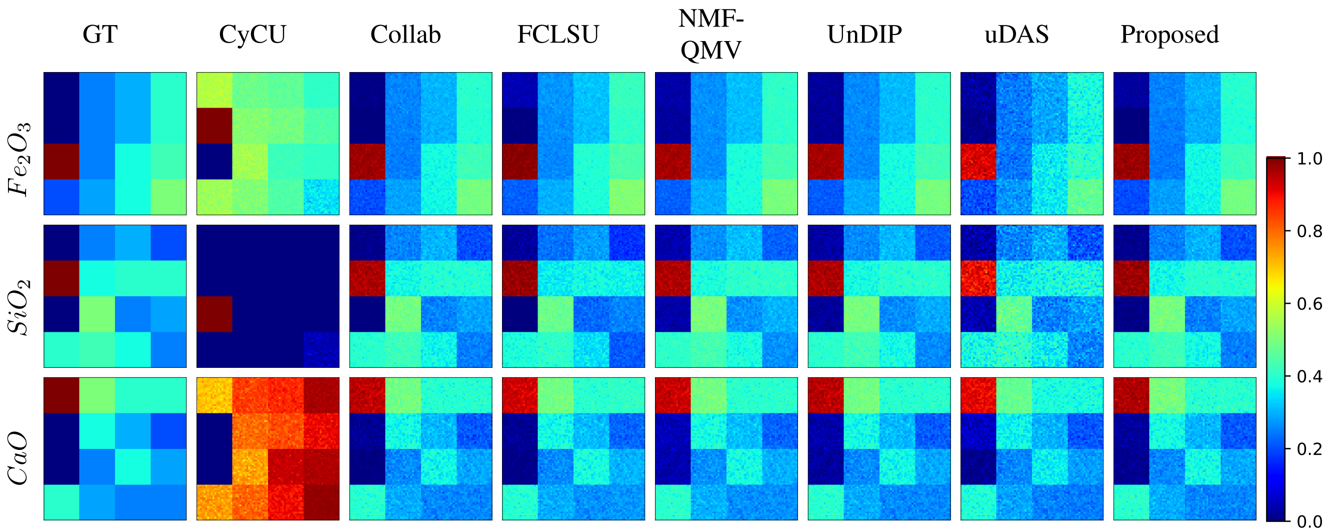


Fig. 7: Simulated dataset (SNR 30dB) - Visual comparison of the abundance maps obtained by the different unmixing techniques.

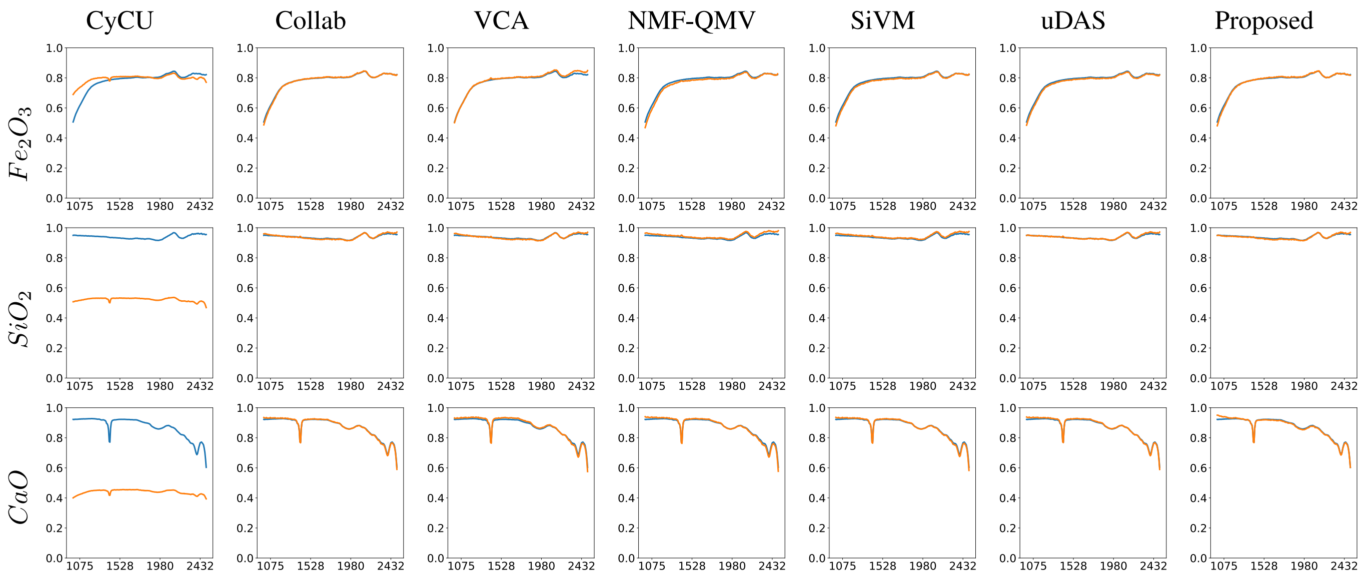


Fig. 8: Simulated dataset - Visual comparison of the endmembers obtained by the different unmixing techniques. Blue: ground truth endmembers; Orange: estimated endmembers.

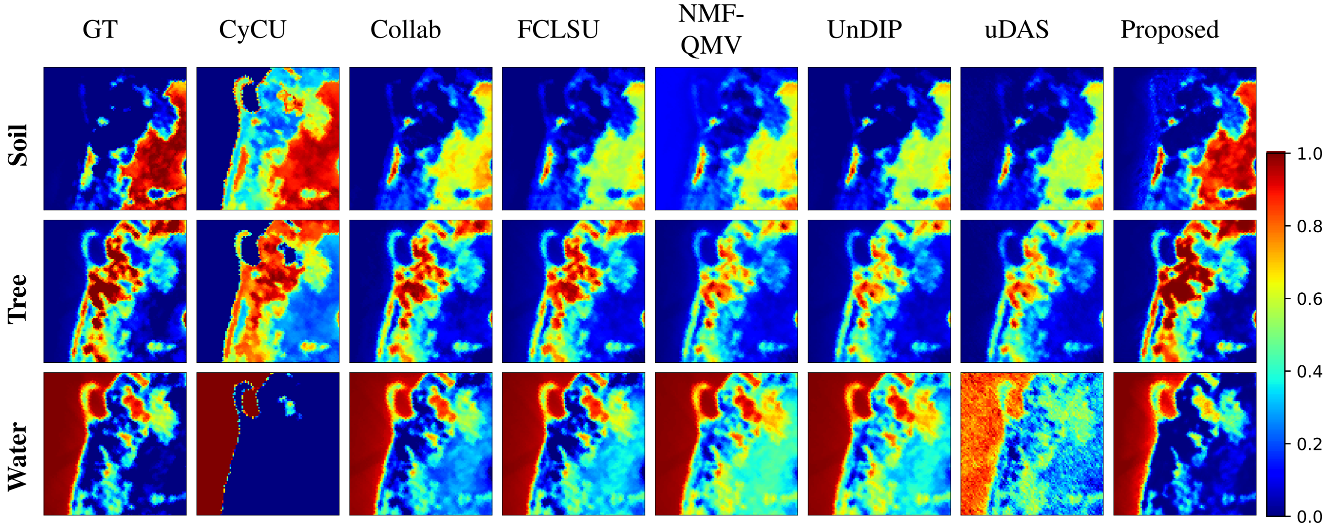


Fig. 9: Samson dataset - Visual comparison of the abundance maps obtained by the different unmixing techniques.

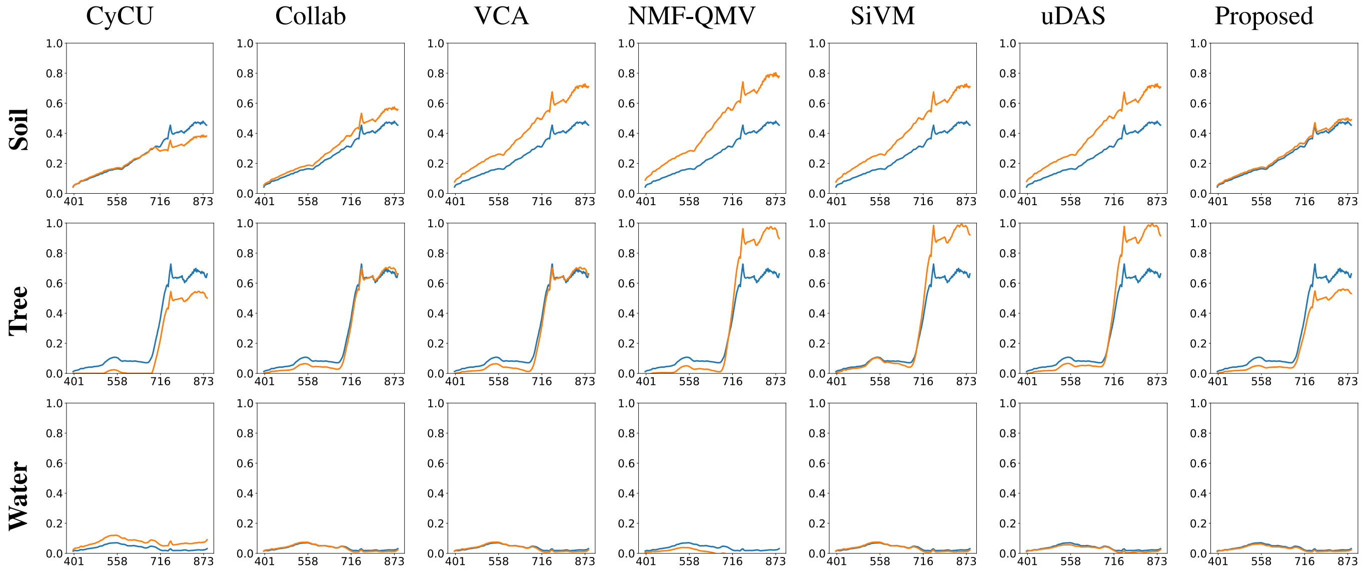


Fig. 10: Samson dataset - Visual comparison of the endmembers obtained by the different unmixing techniques. Blue: ground truth endmembers; Orange: estimated endmembers.

TABLE V: RMSE (SAMSON DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	FCLSU	NMF-QMV	UnDIP	uDAS	Proposed
Soil	0.2417	0.1506	0.1766	0.2011	0.1778	0.1799	**0.0712**
Tree	0.1386	**0.0607**	0.0653	0.1466	0.1330	0.1383	0.0683
Water	0.2654	0.1181	0.1492	0.2063	0.2096	0.2303	**0.0930**
Overall	0.2222	0.1159	0.1387	0.1866	0.1763	0.1867	**0.0783**

TABLE VI: SAD (SAMSON DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	VCA	NMF-QMV	SiVM	uDAS	Proposed
Soil	0.1144	0.0155	0.0259	0.0391	0.0259	0.0358	**0.0128**
Tree	0.1517	0.0832	0.0961	0.1239	0.0748	0.0960	**0.0674**
Water	0.2081	0.1402	0.1554	1.5201	0.1554	0.1527	**0.0729**
Overall	0.1581	0.0796	0.0925	0.5610	0.0854	0.0948	**0.0510**

but modifies the spectral signatures in a way that they more closely resemble the ground truth endmembers, with much lower SAD errors.

To test the effectiveness of the proposed method for estimating endmembers and abundance maps of complex datasets, a cropped image of the Cuprite dataset was considered in this work. This dataset contains 250 × 190 hyperspectral pixels (see Fig. 15(a) for true-color image and Fig. 15(b) for ground truth mineral map). Each hyperspectral pixel contains reflection values for 185 different bands covering the wavelength range [389-2442] nm. The major advantage of this dataset is that it provides a ground-measured spectral library for

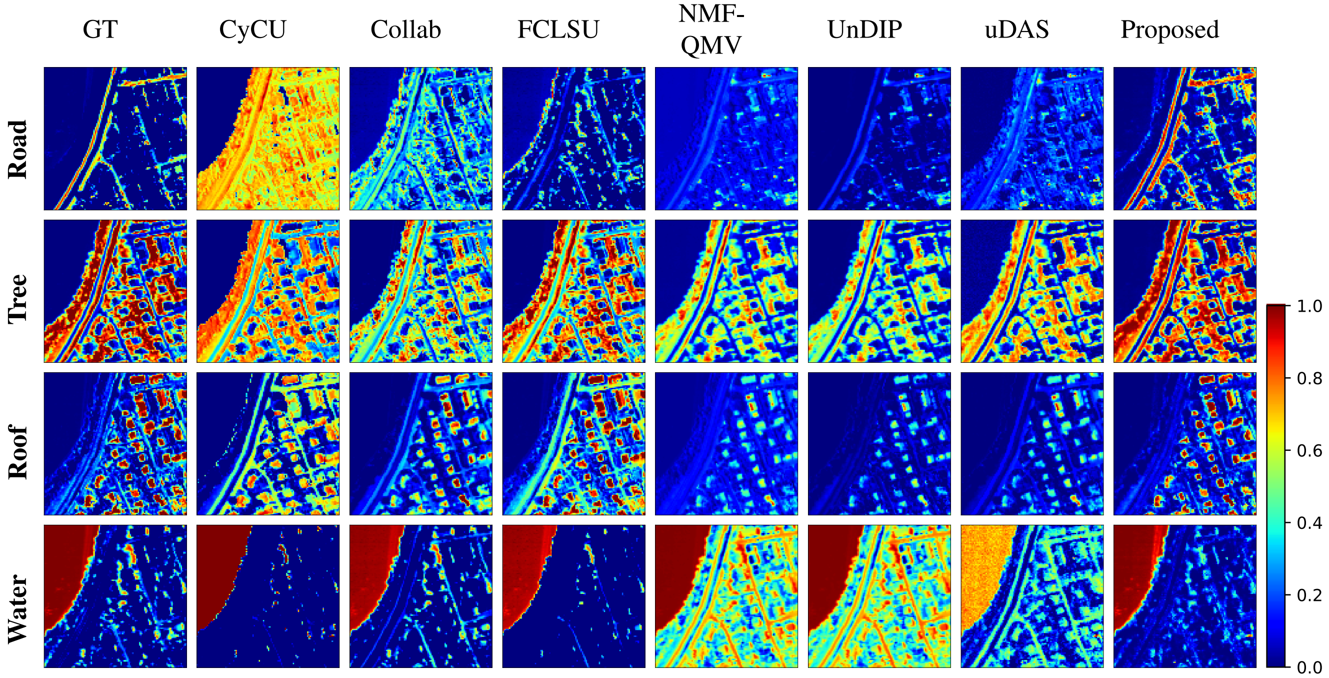


Fig. 11: Apex dataset - Visual comparison of the abundance maps obtained by the different unmixing techniques.

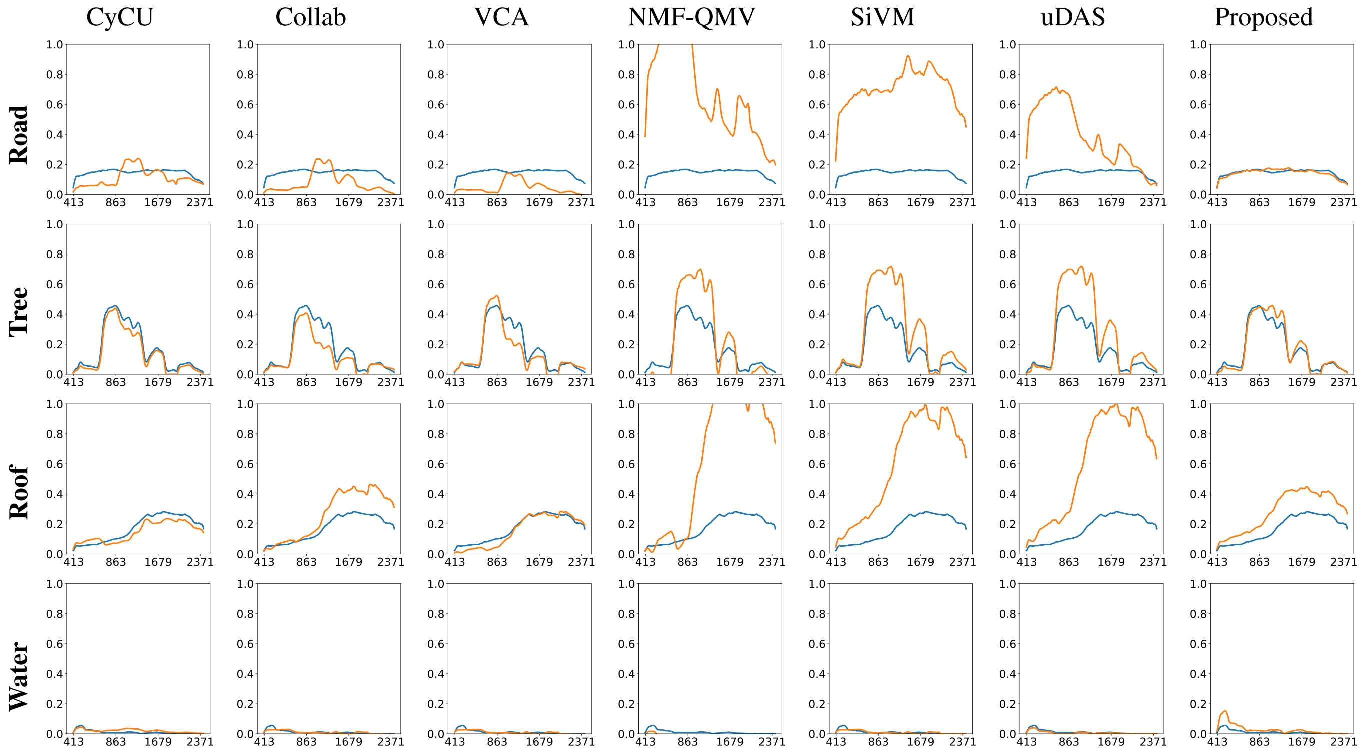


Fig. 12: Apex dataset - Visual comparison of the endmembers obtained by the different unmixing techniques. Blue: ground truth endmembers; Orange: estimated endmembers.

evaluating the estimated endmembers. In this scene, there are 12 materials. Among them, Alunite, Kaolinite1, Kaolinite2, Muscovite, Montmorillonite, Sphene, and Chalcedony are the most dominant ones. The major challenge for this dataset is the production of the ground truth abundance maps for comparison. To prepare realistic ground truth abundance maps, we manually picked spectra of these seven minerals from the image. For this, the ground truth mineral map (see Fig. 15(b)) was used as prior information. The fractional abundances were produced by applying FCLSU. Visually, the obtained

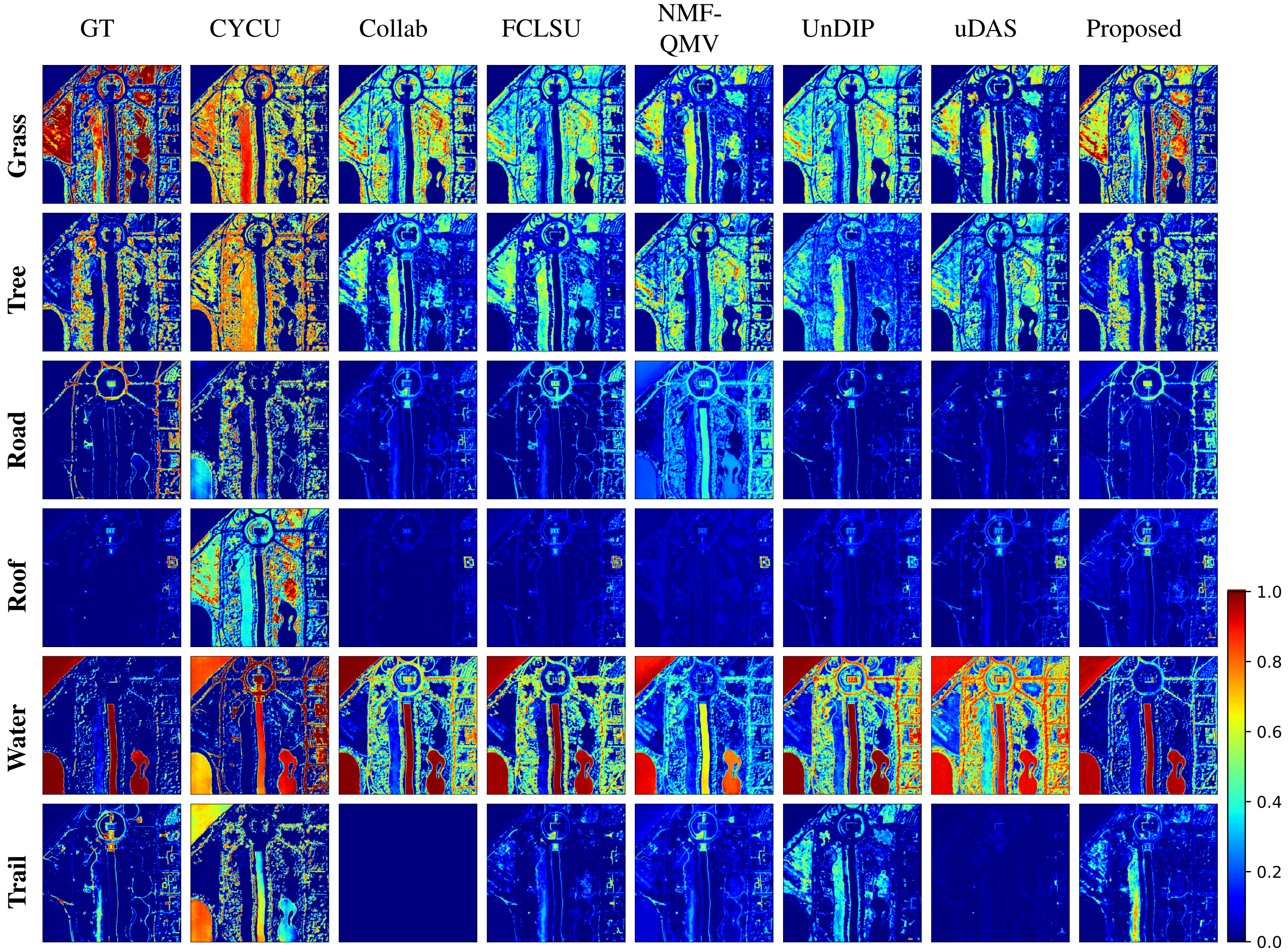


Fig. 13: Washington DC Mall dataset - Visual comparison of the abundance maps obtained by the different unmixing techniques.

TABLE VII: RMSE (APEX DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	FCLSU	NMF-QMV	UnDIP	uDAS	Proposed
Road	0.2921	0.3078	0.2331	0.1806	**0.1737**	0.1973	0.1776
Tree	0.2020	0.1907	**0.0944**	0.2468	0.2154	0.1419	0.0993
Roof	0.1630	0.1483	0.1201	0.2359	0.2554	0.2303	**0.1200**
Water	0.1213	**0.0797**	0.1327	0.3751	0.4170	0.2887	0.0902
Overall	0.2046	0.1997	0.1543	0.2692	0.2809	0.2210	**0.1264**

TABLE VIII: SAD (APEX DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	VCA	NMF-QMV	SiVM	uDAS	Proposed
Road	0.4543	0.6772	0.6915	0.4003	0.0907	0.4551	**0.0836**
Tree	**0.0850**	0.2063	0.2644	0.2710	0.1339	0.1405	0.1295
Roof	0.1298	0.1002	0.1471	0.1753	**0.0689**	0.0860	0.0903
Water	0.6223	0.5137	0.5176	1.8417	0.5040	0.2251	**0.0434**
Overall	0.3228	0.3744	0.4052	0.6721	0.1994	0.2267	**0.0867**

TABLE IX: RMSE (WASHINGTON DC MALL DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CYCU	Collab	FCLSU	NMF-QMV	UnDIP	uDAS	Proposed
Grass	0.4104	0.2901	0.3090	0.3624	0.2978	0.3780	**0.1661**
Tree	0.2824	0.4167	0.4025	0.2761	0.3514	0.3351	**0.0963**
Road	0.2545	0.2263	0.1757	0.2351	0.2436	0.2497	**0.1353**
Roof	0.4157	0.0437	**0.0380**	0.0862	0.0493	0.0463	0.0863
Water	0.3957	0.3102	0.2921	0.2076	0.3812	0.5156	**0.1326**
Trail	0.2072	0.1875	0.1230	**0.1011**	0.2360	0.1769	0.1492
Overall	0.3379	0.2715	0.2550	0.2322	0.2814	0.3206	**0.1307**

TABLE X: SAD (WASHINGTON DC MALL DATASET). THE BEST PERFORMANCES ARE SHOWN IN BOLD.

	CyCU	Collab	VCA	NMF-QMV	SiVM	uDAS	Proposed
Grass	**0.0895**	0.3171	0.3170	0.1952	0.1851	0.1897	0.2379
Tree	0.2704	0.3335	0.2883	0.4507	0.7258	0.4251	**0.1225**
Road	0.4642	0.3439	0.2316	0.2243	0.8608	0.6585	**0.0781**
Roof	0.9500	**0.0331**	0.0343	0.2078	0.2826	0.1992	0.3352
Water	0.4205	**0.0305**	0.7766	0.6736	0.9495	0.2328	0.0533
Trail	0.7906	0.3446	0.6472	**0.0615**	0.1754	0.0941	0.0951
Overall	0.4975	0.2338	0.3825	0.3022	0.5299	0.2999	**0.1537**

ground truth abundance maps represent the scene (see the first row of Fig. 16). It can be observed from the second row of Fig. 16 that the proposed model performed decently

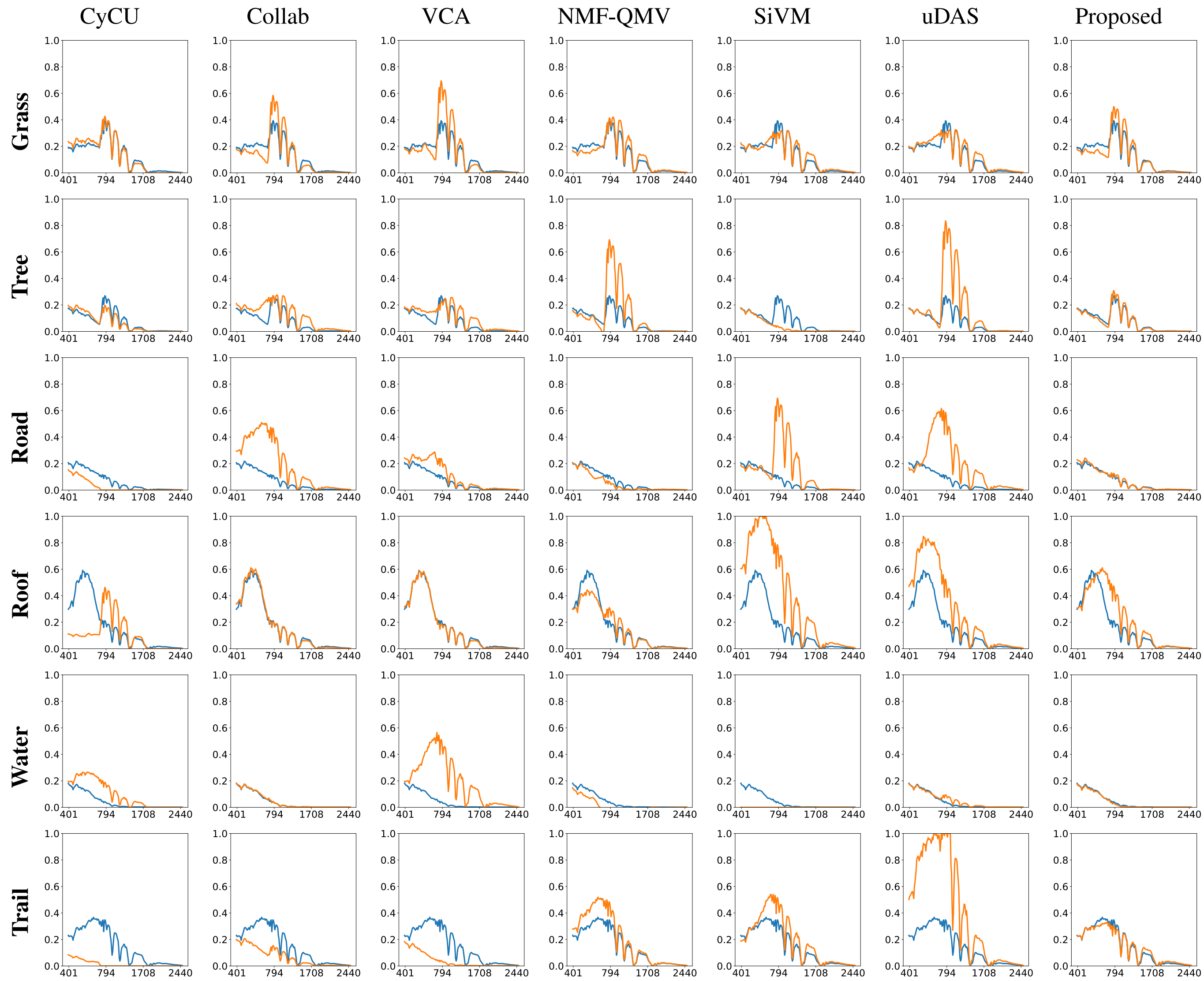


Fig. 14: Washington DC Mall dataset - Visual comparison of the endmembers obtained by the different unmixing techniques. Blue: ground truth endmembers; Orange: estimated endmembers.

on this complex dataset. From Fig. 17, one can observe that the obtained spectral signatures correspond well to the ground truth endmembers, demonstrating that the proposed method can potentially have practical use in real-life scenarios.

G. Sensitivity Analysis to Hyperparameters

The hyperparameters β and γ play essential roles in determining the model's overall performance. In order to keep the training process simple, the value of β was kept constant at 5×10^3 for all real datasets while β was kept constant at 1×10^4 for the simulated dataset. Table XI depicts the sensitivity of the proposed unmixing model to the hyperparameter γ. Changing γ affects both SAD and RMSE similarly in most cases. The table suggests that γ can be set in the range 1×10^{-4} to 1×10^{-2}, with a higher number of endmembers favouring a lower γ value.

Apart from the hyperparameters mentioned above, the learning rate and the weight decay were also found to have a significant impact on the obtained results, as can be observed in Tables XII and XIII. Learning rates were tested in the range from 0.001 to 0.009, and the best results were obtained in the range from 0.006 to 0.009, with images having lower spatial dimensions preferring a slightly lower learning rate. The weight decay was tested in the range from 1×10^{-5} to 9×10^{-5}. Table XIII suggests an optimal value of 4×10^{-5}. It was observed that the quality of the abundance maps quickly deteriorates with increasing weight decay.

The optimal parameters were selected using a grid search-based approach on the sample space [65], and the combination of parameter values which resulted in the minimal value of the loss function in Eq. (12) was finally applied to obtain the reported results.

IV. Conclusion

In this article we proposed a novel HSI unmixing method that uses a convolutional autoencoder combined with a trans-

TABLE XI: DEPENDENCE OF RMSE AND SAD ON GAMMA FOR SIMULATED, SAMSON, APEX AND WASHINGTON DC MALL DATASETS.

Gamma	Simulated		Samson		Apex		WDC Mall	
	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
1×10^{-6}	0.0063	0.0207	0.0751	0.0650	0.2416	0.1444	0.3042	0.2566
5×10^{-6}	0.0063	0.0208	0.1070	0.0532	0.1860	0.1476	0.1910	0.1456
1×10^{-5}	0.0063	0.0208	0.1099	0.0697	0.2538	0.1587	0.2827	0.1835
5×10^{-5}	0.0063	0.0207	0.0797	0.0541	0.2128	0.1486	0.2140	0.1707
1×10^{-4}	**0.0063**	**0.0207**	0.1314	0.0660	0.2185	0.1648	**0.1636**	**0.1355**
5×10^{-4}	0.0063	0.0207	0.1346	0.0805	0.1262	0.1329	0.1778	0.1366
1×10^{-3}	0.0063	0.0208	0.1421	0.0899	0.1192	0.1486	0.1733	0.1274
5×10^{-3}	0.0063	0.0208	**0.0467**	**0.0664**	0.0997	0.1395	0.2681	0.3141
1×10^{-2}	0.0063	0.0208	0.0961	0.0747	0.1268	0.1346	0.2894	0.2678
5×10^{-2}	0.0063	0.0213	0.0760	0.0781	**0.0867**	**0.1264**	0.3016	0.2863
1×10^{-1}	0.0063	0.0221	0.0448	0.1273	0.0818	0.1550	0.3380	0.3180
5×10^{-1}	0.0065	0.0279	0.0430	0.1660	0.1911	0.2024	0.3099	0.2609

TABLE XII: DEPENDENCE OF RMSE AND SAD ON LEARNING RATE WITH FIXED WEIGHT DECAY FOR SIMULATED, SAMSON, APEX AND WASHINGTON DC MALL DATASETS.

Learning Rate	Simulated		Samson		Apex		WDC Mall	
	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
0.001	0.0073	0.0215	0.1215	0.1450	0.2080	0.2884	0.3210	0.2510
0.002	0.0071	0.0214	0.1176	0.1359	0.3001	0.1943	0.4010	0.2496
0.003	0.0067	0.0213	0.1077	0.1254	0.2823	0.1987	0.3762	0.2573
0.004	0.0063	0.0213	0.0975	0.1107	0.2986	0.1948	0.2927	0.1841
0.005	0.0060	0.0213	0.0893	0.0891	0.2857	0.1945	0.1559	0.1160
0.006	0.0056	0.0214	0.0839	0.0770	0.0895	0.1223	0.1525	0.1281
0.007	0.0054	0.0214	**0.0746**	**0.0721**	**0.0951**	**0.1189**	**0.1427**	**0.1141**
0.008	**0.0051**	**0.0212**	0.0271	0.1223	0.2649	0.1396	0.2492	0.3937
0.009	0.0049	0.0215	0.0459	0.2536	0.1096	0.1687	0.1755	0.1528

TABLE XIII: DEPENDENCE OF RMSE AND SAD ON WEIGHT DECAY WITH FIXED LEARNING RATE FOR SIMULATED, SAMSON, APEX AND WASHINGTON DC MALL DATASETS.

Weight Decay	Simulated		Samson		Apex		WDC Mall	
	SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
1×10^{-5}	0.006322	0.021177	0.0491	0.2701	0.1265	0.2247	0.1392	0.2284
2×10^{-5}	0.006323	0.021312	0.0397	0.1012	0.1536	0.2114	0.3488	0.3424
3×10^{-5}	0.006337	0.021317	0.0629	0.0720	0.1474	0.2523	**0.2311**	**0.1643**
4×10^{-5}	0.006341	0.021332	**0.0540**	**0.0963**	**0.0984**	**0.1432**	0.3492	0.1867
5×10^{-5}	**0.006317**	**0.021345**	0.1019	0.0883	0.1194	0.1185	0.3430	0.2308
6×10^{-5}	0.006330	0.021342	0.0914	0.0721	0.1064	0.1269	0.4160	0.1822
7×10^{-5}	0.006333	0.021325	0.1172	0.0987	0.1458	0.1469	0.4490	0.2631
8×10^{-5}	0.006332	0.021149	0.0977	0.0919	0.1359	0.1461	0.4085	0.2423
9×10^{-5}	0.006333	0.021276	0.0899	0.0901	0.1200	0.1263	0.4090	0.1914

former. To the best of our knowledge, this is the first transformer model addressing the hyperspectral unmixing problem. We demonstrated the viability of the novel Multihead Self-Patch Attention mechanism used in the encoder block of the transformer. The experiments were carried out on three real datasets, each with its unique set of challenges, and were successfully handled by the proposed model with consistent performance across the range of endmembers. The accuracy and consistency of the proposed model can be credited to the use of the transformer block which captures the long range feature dependencies that are otherwise not reachable by a CNN based architecture. This enables our model to achieve superior unmixing results, which are significantly better than the competing methods.

Acknowledgment

The authors thank Ganesan Narayanasamy who is leading IBM OpenPOWER/POWER enablement and ecosystem worldwide for his support to get the IBM AC922 system's access. B. Koirala is funded by the Research Foundation-Flanders - Project G031921N.

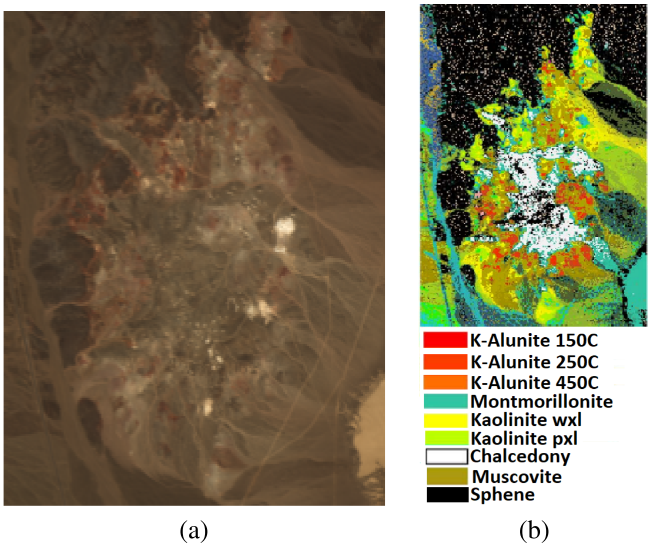
References

[1] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2020.

[2] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 60–88, 2020.

[3] S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza, "Morphological convolutional neural networks for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8689–8702, 2021.

[4] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *arXiv preprint arXiv:2203.16952*, 2022.



(a)

(b)

Fig. 15: Cuprite image: (a) True-color image (Red: 654 nm, Green: 550 nm, Blue: 455 nm); (b) Ground truth mineral map.

[5] B. Koetz, F. Morsdorf, S. Van der Linden, T. Curt, and B. Allgöwer, "Multi-source land cover classification for forest fire management based on imaging spectrometry and lidar data," *Forest Ecology and Management*, vol. 256, no. 3, pp. 263–271, 2008.

[6] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 968–999, 2022.

[7] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognition*, vol. 48, no. 12, pp. 3904–3916, 2015.

[8] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[9] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.

[10] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, April 2012.

[11] D. C. Heinz and Chein-I-Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.

[12] R. Heylen, D. Burazerovic, and P. Scheunders, "Fully constrained least squares spectral unmixing by simplex projection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4112–4122, Nov 2011.

[13] J. Nascimento and J. Bioucas-Dias, "Vertex component analysis: A˜fast algorithm to extract endmembers spectra from hyperspectral data," in *Pattern Recognition and Image Analysis*, F. J. Perales, A. J. C. Campilho, N. P. de la Blanca, and A. Sanfeliu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 626–635.

[14] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4418–4432, 2009.

[15] M. E. Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *Imaging Spectrometry V*, M. R. Descour and S. S. Shen, Eds., vol. 3753, International Society for Optics and Photonics. SPIE, 1999, pp. 266 – 275. [Online]. Available: https://doi.org/10.1117/12.366289

[16] W. Full, R. Ehrlich, and J. Klovan, "Extended qmodel—objective definition of external end members in the analysis of mixtures," *Journal of the International Association for Mathematical Geology*, vol. 13, pp. 331–344, 08 1981.

[17] M. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552, 1994.

[18] E. M. T. Hendrix, I. Garcia, J. Plaza, G. Martin, and A. Plaza, "A new minimum-volume enclosing algorithm for endmember identification and abundance estimation in hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2744–2757, 2012.

[19] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, 2008, pp. III – 250–III – 253.

[20] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009, pp. 1–4.

[21] N. Dobigeon, S. Moussaoui, M. Coulon, J. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

[22] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.

[23] J. Li, J. M. Bioucas-Dias, A. Plaza, and L. Liu, "Robust collaborative nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6076–6090, 2016.

[24] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Collaborative nonnegative matrix factorization for remotely sensed hyperspectral unmixing," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2012, pp. 3078–3081.

[25] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. Huntington, "Ice: a statistical approach to identifying endmembers in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 10, pp. 2085–2095, 2004.

[26] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization for remote sensing and document clustering," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6254–6268, 2016.

[27] X. Xu, J. Li, S. Li, and A. Plaza, "Curvelet transform domain-based sparse nonnegative matrix factorization for hyperspectral unmixing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4908–4924, 2020.

[28] C. Chenot and J. Bobin, "Blind source separation with outliers in transformed domains," *SIAM Journal on Imaging Sciences*, vol. 11, no. 2, pp. 1524–1559, 2018.

[29] V. S S and J. S. Bhatt, "A blind spectral unmixing in wavelet domain," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 287–10 302, 2021.

[30] L. Zhuang, C. Lin, M. A. T. Figueiredo, and J. M. Bioucas-Dias, "Regularization parameter selection in minimum volume hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9858–9877, 2019.

[31] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.

[32] B. Rasti, B. Koirala, P. Scheunders, and J. Chanussot, "Misicnet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[33] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

[34] J. M. P. Nascimento and J. M. Bioucas-Dias, "Hyperspectral unmixing based on mixtures of dirichlet components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 863–878, 2012.

[35] M. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2014–2039, 2011.

[36] ——, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4484–4502, 2012.

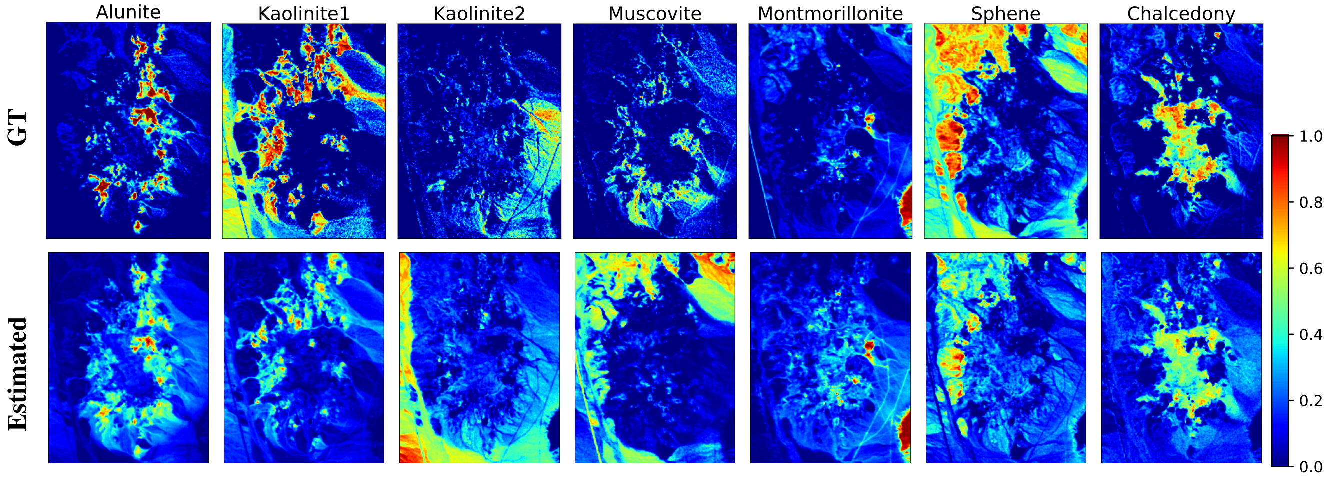


Fig. 16: Cuprite dataset - Visual comparison of the estimated abundance maps with the proposed method and the ground truth abundance maps.

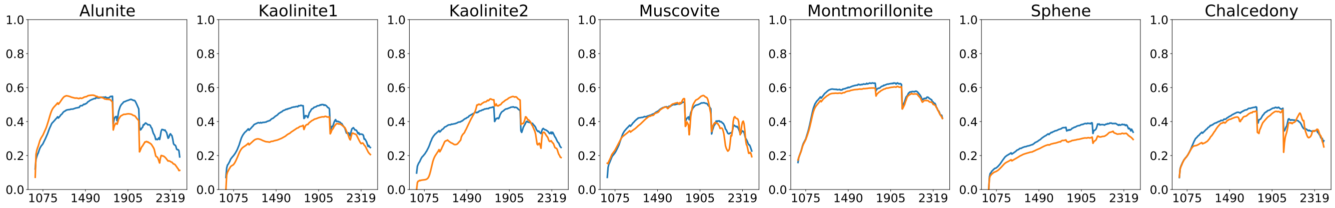


Fig. 17: Cuprite dataset - Visual comparison of the extracted and ground truth endmembers. Blue: ground truth endmembers; Orange: estimated endmembers by the proposed method.

[37] B. Rasti and B. Koirala, "Suncnn: Sparse unmixing using unsupervised convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[38] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[39] S. K. Roy, P. Kar, D. Hong, X. Wu, A. Plaza, and J. Chanussot, "Revisiting deep hyperspectral feature extraction networks via gradient centralized convolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[40] J. S. Bhatt and M. V. Joshi, "Deep learning in hyperspectral unmixing: A review," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2189–2192.

[41] J. R. Patel, M. V. Joshi, and J. S. Bhatt, "Spectral unmixing using autoencoder with spatial and spectral regularizations," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3321–3324.

[42] V. S. S, V. S. Deshpande, and J. S. Bhatt, "A practical approach for hyperspectral unmixing using deep learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[43] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–1, 2022.

[44] Y. Su, A. Marinoni, J. Li, J. Plaza, and P. Gamba, "Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1427–1431, 2018.

[45] Y. Su, J. Li, A. Plaza, A. Marinoni, P. Gamba, and S. Chakravortty, "Daen: Deep autoencoder networks for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4309–4321, 2019.

[46] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Deep generative endmember modeling: An application to unsupervised spectral unmixing," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 374–384, 2020.

[47] Q. Jin, Y. Ma, F. Fan, J. Huang, X. Mei, and J. Ma, "Adversarial autoencoder network for hyperspectral unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[48] M. Tang, Y. Qu, and H. Qi, "Hyperspectral nonlinear unmixing via generative adversarial network," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 2404–2407.

[49] S. K. Roy, J. M. Haut, M. E. Paoletti, S. R. Dubey, and A. Plaza, "Generative adversarial minority oversampling for spectral–spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[50] Y. Qu and H. Qi, "udas: An untied denoising autoencoder with sparsity for spectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1698–1712, 2019.

[51] X. Zhang, Y. Sun, J. Zhang, P. Wu, and L. Jiao, "Hyperspectral unmixing via deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1755–1759, 2018.

[52] B. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Convolutional autoencoder for spectral-spatial hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2020.

[53] L. Gao, Z. Han, D. Hong, B. Zhang, and J. Chanussot, "Cycu-net: Cycle-consistency unmixing network by learning cascaded autoencoders," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021.

[54] D. Hong, L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, and B. Zhang, "Endmember-guided unmixing network (egu-net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021.

[55] Y. Su, X. Xu, J. Li, H. Qi, P. Gamba, and A. Plaza, "Deep autoencoders with multitask learning for bilinear hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8615–8629, 2021.

[56] F. Khajehrayeni and H. Ghassemian, "Hyperspectral unmixing using deep convolutional autoencoders in a supervised scenario," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 567–576, 2020.

[57] B. Rasti, B. Koirala, P. Scheunders, and P. Ghamisi, "UnDIP: Hyperspectral unmixing using deep image prior," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021.

[58] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 357–366.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[60] J. Guo, K. Han, H. Wu, C. Xu, Y. Tang, C. Xu, and Y. Wang, "Cmt: Convolutional neural networks meet vision transformers," *arXiv preprint arXiv:2107.06263*, 2021.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[62] F. Zhu, Y. Wang, B. Fan, S. Xiang, G. Meng, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5412–5427, 2014.

[63] M. E. Schaepman, M. Jehle, A. Hueni, P. D'Odorico, A. Damm, J. Weyermann, F. D. Schneider, V. Laurent, C. Popp, F. C. Seidel, K. Lenhard, P. Gege, C. Küchler, J. Brazile, P. Kohler, L. De Vos, K. Meuleman, R. Meynart, D. Schläpfer, M. Kneubühler, and K. I. Itten, "Advanced radiometry measurements and earth science applications with the airborne prism experiment (apex)," *Remote Sensing of Environment*, vol. 158, pp. 207–219, 2015.

[64] L. Drumetz, T. R. Meyer, J. Chanussot, A. L. Bertozzi, and C. Jutten, "Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3435–3450, 2019.

[65] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.

Preetam Ghosh completed his bachelor's degree in Computer Science and Engineering in the year 2022 from Jalpaiguri Government Engineering College, Jalpaiguri. He is currently working as a Systems Engineer in Tata Consultancy Services Ltd. His future plan is to pursue a career as a data science engineer. His research interest includes computer vision, deep learning and remote sensing.

Swalpa Kumar Roy (S'15) received the bachelor's and the master's degree in Computer Science and Engineering from West Bengal University of Technology, Kolkata, India, in 2012, and Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India in 2015 and also the Ph.D. degree in Computer Science and Engineering from University of Calcutta, Kolkata, India in 2021.

From July 2015 to March 2016, he was a Project Linked Person with the Optical Character Recognition (OCR) Laboratory, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, West Bengal, India. Dr. Roy was nominated for the Indian National Academy of Engineering (INAE) engineering teachers mentoring fellowship program by INAE Fellows in academic tenure 2021-22 and also a recipient of the Outstanding Paper Award in second Hyperspectral Sensing Meets Machine Learning and Pattern Analysis (HyperMLPA) at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021. He has served as a reviewer for the IEEE TGRS and IEEE GRSL. His research interests include computer vision, deep learning and remote sensing.

Bikram Koirala (Member, IEEE) received the M.S. degree in geomatics engineering from the University of Stuttgart, Stuttgart, Germany, in 2016, and the Ph.D. degree in development of advanced hyperspectral unmixing methods from the University of Antwerp, Antwerp, Belgium, in 2021.

In 2017, he joined the Vision Lab, Department of Physics, University of Antwerp, as a Ph.D. Researcher, where he is currently a Post-Doctoral Researcher. His research interests include machine learning and hyperspectral image processing.

Behnood Rasti (M'12–SM'19) received the B.Sc. and M.Sc. degrees both in electronics- electrical engineering from the Electrical Engineer- ing Department, University of Guilan, Rasht, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2014. In 2015 and 2016, he worked as a Post-Doctoral Researcher with Electrical and Computer Engineering Department, University of Iceland. From 2016 to 2019, he has been a Lecturer with the Center of Engineering Technology and Applied Sciences, Department of Electrical and Computer Engineering, University of Iceland. Dr. Rasti was a Humboldt research fellow in 2020 and 2021. He is currently a Principal Research Associate with Helmholtz-Zentrum Dresden-Rossendorf (HZDR). His research interests include machine/deep learning, signal and image processing, remote sensing, and artificial intelligence.

Dr. Rasti was the Valedictorian as an M.Sc. Student in 2009. He won the Doctoral Grant of The University of Iceland Research Fund "The Eimskip University fund," and the "Alexander von Humboldt Research Fellowship Grant" in 2013 and 2019, respectively. He serves as an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL).

Paul Scheunders (M'98) received the B.S. degree and the Ph.D. degree in physics, with work in the field of statistical mechanics, from the University of Antwerp, Antwerp, Belgium, in 1983 and 1990, respectively. In 1991, he became a research associate with the Vision Lab, Department of Physics, University of Antwerp, where he is currently a full professor. His current research interest includes remote sensing and hyperspectral image processing. He has published over 200 papers in international journals and proceedings in the field of image processing, pattern recognition, and remote sensing. Paul Scheunders is Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing and has served as a program committee member in numerous international conferences. He is a senior member of the IEEE Geoscience and Remote Sensing Society.